Diebold, F.X. (2003),
"'Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting"
(Discussion of Reichlin and Watson papers), in M. Dewatripont, L.P. Hansen and S.Turnovsky (Eds.),
Advances in Economics and Econometrics, Eighth World Congress of the Econometric Society.
Cambridge: Cambridge University Press, 115-122.

"Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting

Francis X. Diebold

University of Pennsylvania
and NBER

First Version, July 2000
November 28, 2000

Correspondence to:

F.X. Diebold
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297

Email: fdiebold@mail.sas.upenn.edu
Home Page: http://www.ssc.upenn.edu/~diebold
Phone: (610) 585-4057
Fax: (215) 573-4217

This note is a discussion of "Extracting Business Cycle Indexes from Large Data Sets: Aggregation, Estimation, Identification," by Lucrezia Reichlin, and "Macroeconomic Forecasting Using Many Predictors," by Mark Watson, prepared for the World Congress of the Econometric Society, Seattle, August 2000. I am grateful to the National Science Foundation for its support, and to Rob Engle, Jim Stock and Mark Watson for helpful comments.

1. "Big Data"

The Reichlin and Watson papers are just what we've come to expect from their authors: practical and pragmatic, yet grounded in rigorous theory. In short, good science.

Recently much good science, whether physical, biological, or social, has been forced to confront – and has often benefitted from – the "Big Data" phenomenon. Big Data refers to the explosion in the quantity (and sometimes, quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology. In this new and exciting world, sample sizes are no longer fruitfully measured in "number of observations," but rather in, say, megabytes. Even data accruing at the rate of several gigabytes per day are not uncommon. Economics examples include microeconometric analyses of consumer choice, which have been transformed by the availability of huge and detailed datasets collected by checkout scanners, and financial econometric analyses of asset return dynamics, which have been similarly transformed by the availability of tick-by-tick data for thousands of assets.

From the vantage point of the examples sketched above, the Reichlin and Watson papers don't analyze *really* Big Data, but they certainly represent a movement of macroeconometrics in that direction. In traditional small-data macroeconometric environments such as Stock and Watson (1989), one might work with a vector autoregression involving, say, four or five macroeconomic indicators measured over 150 quarters. In contrast, Reichlin and Watson work with roughly 500 indicators measured over 500 months. In the not-too-distant future, we will be working with thousands of indicators, with many measured at daily or higher frequency.

A canonical problem in the analysis of Big Data is how to make tractable an "X matrix" (in regression parlance) of dimension TxK, when both T and K are very large. In the applications at hand, the variables in X are used to extract or forecast the state of macroeconomic activity (f), on which they depend. The latent factor f may be the object of intrinsic interest (Reichlin), or it may be used to forecast some other variable, y=g(f) (Watson). Many approaches proceed by somehow reducing the number of columns of X. Variable selection methods, for example, whether based on the traditional tools of inference such as t and F tests or on more recently-developed criteria such as AIC, amount to strategies for eliminating columns of X. The principal component methods used by Reichlin and Watson are rather more sophisticated, not requiring a sharp "in" or "out" decision for each variable, but rather allowing all variables to contribute to an extraction or forecast. Of course, replacing a large set of variables with a small set of their first few principal components can not in general be done without substantial information loss. It is truly fortunate for macroeconomists and financial economists that our data are often well-approximated by low- dimensional factor structures, in which case replacing large sets of variables by a few principal components is not only convenient, but also legitimate (in the sense of little information loss).

2. Factor Structure and Regime Switching

Let us elaborate on the idea of factor structure. To do so it will be helpful to recall what I call the linear tradition in macroeconomics and business cycle analysis, which emphasizes the modeling and interpretation of comovement among macroeconomic aggregates, as in Burns and Mitchell (1946). Part of the modern econometric distillation of that tradition is the vector autoregression, a linear model that captures comovement by allowing lags of variable i to affect variable j, for all i and j, and by allowing for contemporaneous correlation across shocks. But the vector autoregression alone does not provide a viable description of large sets of macroeconomic indicators; degrees of freedom would soon be

exhausted. Hence the appeal of dynamic factor structure (Sargent and Sims, 1977; Geweke, 1977), reflecting the recognition that comovements among macroeconomic indicators likely arise from partial dependence on common shocks. In a one-factor model, which is the leading case in practice, there is just one common shock. Hence the behavior of each of a potentially large set of K variables is qualitatively similar to the behavior of just one variable, the common factor.

There is also, however, a distinct nonlinear tradition in macroeconomics and business cycle analysis, which emphasizes the idea of regime switching, namely that expansions and contractions may be usefully treated as different probabilistic objects, with turning points naturally defined as switching times. This view is clearly delineated in classics such as Burns and Mitchell (1946) and is embodied in modern regime-switching models, in particular the popular Markov switching model of Hamilton (1989). In Hamilton's model, conditional densities are governed by a parameter vector whose value depends on the state (expansion or contraction, say), with state transitions governed by a first-order Markov process.

The linear and nonlinear traditions have matured largely in isolation, but they are in no way contradictory, and accurate business cycle measurement and forecasting may require elements of both. Hence Diebold and Rudebusch (1996) propose a dynamic factor model in which the factor may display Markov switching. For concreteness, and because it will feature prominently in the sequel, I will sketch a simple one-factor model with first-order autoregressive dynamics and a mean growth rate that switches across expansions and contractions. In an obvious notation,

$$x_t = \beta + \lambda f_t + u_t, \tag{1}$$

where x is a vector of covariance stationary indicators, beta is a vector of constants, λ is a vector of factor loadings, f is a scalar latent common factor, and u is a vector of idiosyncratic shocks. The conditional density of the common factor f is assumed Gaussian, with first-order autoregressive dynamics and a switching unconditional mean,

$$P(f_t|h_t;\theta) \propto \exp\left(\frac{-1}{2\sigma^2}[(f_t-\mu_{s_t}) - \rho(f_{t-1}-\mu_{s_{t-1}})]^2 \right), \tag{2}$$

where h contains past x and f. Finally, the latent state takes the value s=0 in expansions and s=1 in recessions; its dynamics are first-order Markov, with transition dynamics

$$M = \begin{pmatrix} p_{00} & 1-p_{00} \\ 1-p_{11} & p_{11} \end{pmatrix}. \tag{3}$$

Generalizations could be entertained, such as incorporating time-varying transition probabilities as in Diebold, Lee and Weinbach (1994), but the simple model (1) - (3) is well-suited to our present needs.

3. The Evidence

The "small data" variant of the regime-switching dynamic-factor model has met with empirical success. Diebold and Rudebusch (1996) show that, although evidence of regime switching is hard to uncover in the four individual indicators the comprise the coincident index (i.e., common factor) extracted by Stock and Watson (1989), there is strong evidence of regime switching in the index itself. This is precisely what one expects in a regime-switching dynamic-factor world, because the various individual indicators are contaminated by idiosyncratic noise, whereas the common factor is not.

The Diebold-Rudebusch (1996) approach is based on a two-step analysis in which one first extracts a small-data common factor using the Kalman filter and then fits a Markov switching model to the extracted factor. Simultaneous one-step estimation is difficult, due to the two levels of latency in the model: the common factor is latent, and the conditional dynamics of the factor are themselves dependent on a latent state. In important recent work, however, Kim and Nelson (1998, 1999) have used Markov chain Monte Carlo methods to perform one-step maximum likelihood (as well as Bayesian) estimation of the model, confirming and extending the earlier results.

To illustrate and confirm the evidence for regime switching in small-data dynamic factor models for macroeconomic analysis, I fit the Markov-switching model (1) - (3) to a common factor extracted from a very small number of indicators using the Stock-Watson (1989) methodology (i.e., their coincident index, *XCI*, obtained from the NBER's web page) . The sample period is 1959:4 - 1998:12, and the variable modeled is $\Delta \ln XCI$, standardized to have zero mean and unit variance. The results appear in Table 1 and seem to clearly indicate switching across positive and negative growth regimes. The corresponding extracted recession probabilities, shown in Figure 1a, show remarkable conformity to the shaded NBER recession chronology. The likelihood ratio test statistic for the null hypothesis of one state against the alternative of two is a large 29.5; despite the fact that it does not have a chi square distribution (because of the non-identification of nuisance parameters under the null, among other things), the value of 29.5 would likely remain highly significant even if Hansen's (1996) methods were used to generate corrected critical values.

Now let's get back to Big Data. Call the first principal component extracted from Watson's Big Data *PC1*. It turns out that movements in *PC1* and $\Delta \ln XCI$ (both standardized to have zero mean and unit variance) cohere very closely, as evidenced in the scatterplot in Figure 2 and the time series plot in Figure 3. But if *PC1* is very close to $\Delta \ln XCI$, and $\Delta \ln XCI$ is well approximated by a Markov switching process, then should not *PC1* be as well? In Table 2 I show the results of fitting a Markov switching model to *PC1*, and in Figure 1b I show the corresponding extracted recession probabilities. The evidence seems strongly in favor of switching; the likelihood ratio test statistic is 33.7.

4. Conclusions and Directions for Future Research

The Reichlin and Watson papers, and the emerging Big Data dynamic factor modeling literature of which they are an important part, are major contributions to empirical macroeconomics. They are, however, based on linear models and perspectives, whereas even the quick analyses performed here reveal a potentially important nonlinearity – regime switching – lurking just below the surface. Additional research in that direction will likely be fruitful.

Note well that regime switching does not necessarily invalidate the Big Data factor extractions of Reichlin and of Watson. Under conditions, their extractions are consistent for the factor even with regime switching (a fact which, by the way, could presumably be used to develop formal justification for the Diebold-Rudebusch two-step procedure). Hence regime switching needn't be of central importance if, like Watson, one wants to use f to forecast y, because the factor is still extracted consistently and E(y|f) may be linear even if the dynamics of f are nonlinear. But regime switching, if present, seems more unavoidably central in analyses like Reichlin's, in which interest centers on the monitoring, interpreting and forecasting f.

The upshot is simply that, one way or another, the macroeconometric Big Data dynamic factor modeling literature should think harder about nonlinearity in general, and regime switching in particular. Allowing for regime switching in a dynamic factor model is one way of combining nonlinear with linear dynamics, and recent small data work by Stock and Watson (1999) indicates that there may be forecasting gains from doing so, despite their finding that linear models dominate nonlinear models when one or the other is used alone. The regime switching dynamic factor model is not the only way to combine nonlinear and linear dynamics, and it may not be the best way. But that's really the point: potentially important stones have been left unturned, and much is yet to be learned.

It seems appropriate to close with a forecast. Where is macroeconomic measurement and forecasting based on Big Data dynamic factor models headed? It is headed, and should be headed, toward calculation of recession probabilities conditional on the huge amount of data actually available (many thousands of series), in real-time as the data are released, some measured quarterly, some monthly, some weekly, some such as asset prices in nearly continuous time, and some irregularly. And I conjecture – as is no surprise by now – that the recession probability calculations and forecasts will fruitfully be based on a model with a regime-switching factor. Finally, most of the work in the Big Data factor model literature has been likelihood-based, as has all of this discussion, but I look forward to the formal blending of prior and likelihood information via Bayesian methods, and to assessing the robustness of posterior recession probabilities to alternative prior views. Both Reichlin and Watson are taking important and laudable steps in that direction.

References

Burns, Arthur F., and Wesley C. Mitchell, *Measuring Business Cycles* (New York, New York: National Bureau of Economic Research, 1946).

Diebold, Francis X., Joon-Haeng Lee, and Gretchen C. Weinbach, "Regime Switching with Time-Varying Transition Probabilities," in C. Hargreaves (ed.), *Nonstationary Time- Series Analysis and Cointegration*. Oxford: Oxford University Press, 1994), 283-302. (Reprinted in Diebold, F.X. and G.D. Rudebusch, 1999, *Business Cycles: Durations, Dynamics, and Forecasting*. Princeton: Princeton University Press.)

Diebold, Francis X., and Glenn D. Rudebusch, "Measuring Business Cycles: A Modern Perspective," *Review of Economics and Statistics* 78 (February 1996), 67-77.

Geweke, John, "The Dynamic Factor Analysis of Economic Time-Series Models," in D.J. Aigner and A.S. Goldberger (eds.), *Latent Variables in Socioeconomic Models,* (Amsterdam: North-Holland, 1977), 365-383.

Hamilton, James D., "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica* 57 (March 1989), 357-384.

Hansen, Bruce E., "Inference When a Nuisance Parameter is not Identified Under the Null Hypothesis," *Econometrica* 64 (March 1996), 416-430.

Kim, Chang-Jin and Charles R. Nelson, Business Cycle Turning Points, A New Coincident Index, and Tests of Duration Dependence Based on A Dynamic Factor Model with Regime-Switching, *Review of Economics and Statistics* 80 (May 1998), 188-201.

Kim, Chang-Jin and Charles R. Nelson, *State Space Models with Regime Switching* (Cambridge, Mass., MIT Press, 1999).

Sargent, Thomas J., and Christopher Sims, "Business Cycle Modeling Without Pretending to Have Too Much a Priori Theory," in C. Sims (ed.), *New Methods of Business Cycle Research* (Minneapolis: Federal Reserve Bank of Minneapolis, 1977).

Stock, James H., and Mark W. Watson, "New Indexes of Coincident and Leading Economic Indicators," in O. Blanchard and S. Fischer (eds.), *NBER Macroeconomics Annual* (Cambridge, Mass.: MIT Press, 1989), 351-394.

Stock, James H., and Mark W. Watson, "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series," in R. Engle and H. White (eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W.J. Granger* (Oxford: Oxford University Press, 1999), 1-44.

Table 1
Parameter Estimates
Markov Switching Model, $\Delta \ln XCI$

$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{\rho}$	$\hat{\sigma}$	$\hat{p}_{00}$	$\hat{p}_{11}$
.18	-1.24	.22	.78	.97	.80
(3.9)	(-7.2)	(4.5)	(26.6)	(11.8)	(3.0)

Notes: The table contains maximum likelihood estimates of the parameters of the Markov switching model given by (1)-(3) in the text, with t-statistics in parentheses. The variable modeled is $\Delta \ln XCI$ (standardized to have zero mean and unit variance), where *XCI* is the Stock-Watson (1989) coincident index. The sample period is 1959:4 - 1998:12. See text for details.

Table 2
Parameter Estimates
Markov Switching Model, *PC1*

$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{\rho}$	$\hat{\sigma}$	$\hat{p}_{00}$	$\hat{p}_{11}$
.12	-1.11	.48	.66	.98	.83
(3.2)	(-5.6)	(11.4)	(26.9)	(10.5)	(2.9)

Notes: The table contains maximum likelihood estimates of the parameters of the Markov switching model given by (1)-(3) in the text, with t-statistics in parentheses. The variable modeled is Watson's first principal component, *PC1* (standardized to have zero mean and unit variance). The sample period is 1959:4 - 1998:12. See text for details.

Figure 1a
Smoothed Recession Probabilities, $\Delta \ln XCI$

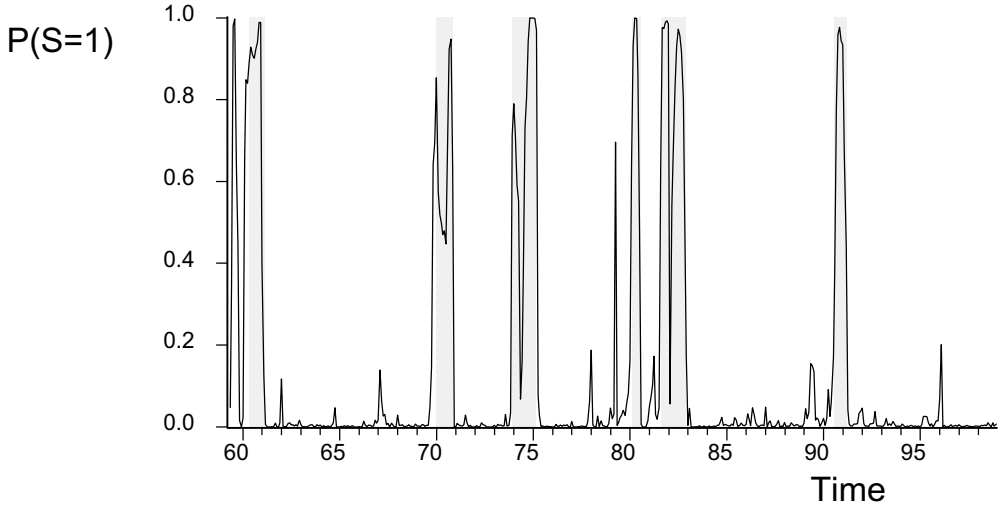

Figure 1b
Smoothed Recession Probabilities, PC1

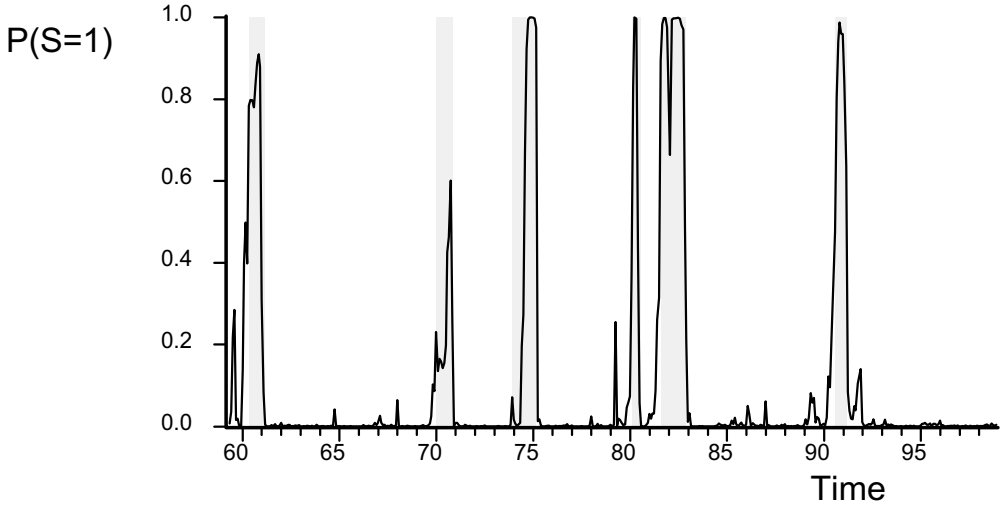

Figure 2
Scatterplot with Fitted Regression Line
PC1 vs. Δln*XCI*

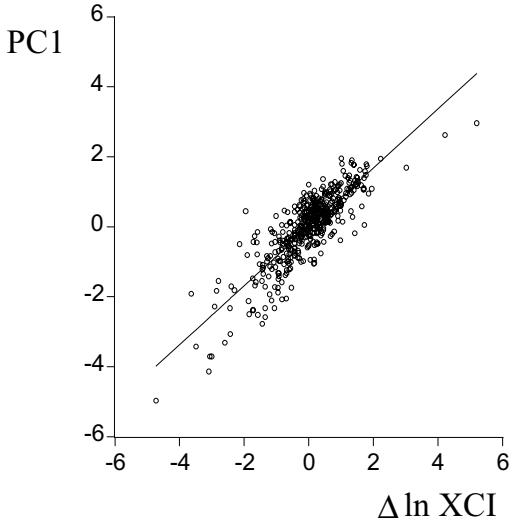

Figure 3
Time Series Plots
PC1 and Δln*XCI*